图书情报工作
第62卷 第10期 2018年5月

基于改进PageRank算法的核心专利发现研究

■ 张欣[1,2] 马瑞敏[3]

[1] 中国科学院文献情报中心 北京100190 [2] 中国科学院大学经济与管理学院 北京100190
[3] 山西大学经济与管理学院 太原030006

摘要：[目的/意义]核心专利的发现是技术创新的重要环节，对于技术改进和专利战略布局意义重大。[方法/过程]首先界定核心专利的概念，然后在对原始PageRank算法模型介绍的基础上，结合专利的被引次数和专利的年龄对原始的PageRank算法进行改进，提出PatentRank算法（简称PTR），并将其应用到OLED领域中来识别核心专利。[结果/结论]研究发现，相比被引次数，PTR不仅能将该领域的核心专利识别出来，而且还可以识别出一些重要性的基础性专利，为研究OLED相关技术提供追本溯源的研究思路；相比原始PageRank算法，PTR具有更高的值区分度，在局部改善了PR的排名。

关键词：PageRank 核心专利 OLED
分类号：G306
DOI：10.13266/j.issn.0252-3116.2018.10.014

1 引言

2008年《国家知识产权战略纲要》中指出：当今世界，随着知识经济和经济全球化深入发展，知识产权日益成为国家发展的战略性资源和国际竞争力的核心要素，成为建设创新型国家的重要支撑和掌握发展主动权的关键[1]。专利作为知识产权的重要组成部分体现了一个国家自主创新的能力。据世界知识产权组织统计，世界上90%－95%的发明都能在专利文献中找到[2-3]，由此可见，专利尤其是核心专利代表着行业内最先进的技术，具有重要的技术价值和市场经济价值。对于高科技型企业来说，站在专利整体布局的高度上，核心专利的发现可以帮助企业围着核心专利进行外围专利的布局或者寻求密切合作伙伴，形成严密的专利网，从而巩固企业的核心竞争力；而对于新兴公司，通过识别当前领域的核心专利并进行技术跟踪，可以减少无谓的研发经费，做到"心中有数"并规避风险。然而，当前每年申请的专利数量庞大，专利的质量也层次不齐。所以，如何挖掘核心专利是摆在我们面前亟待解决的问题。

关于核心专利的概念，目前没有统一的界定，韩志华[4]曾经给出了这样的定义："某一技术领域中处于关键地位、对技术创新具有突出贡献、对其他专利或者技术具有重大影响且具有重要经济价值的专利"。但是，就现在已经建成的公开专利数据资源来看，一个专利的经济价值很难获得精准的数据。所以，本文主要聚焦在"对其他专利具有重大影响"的度量上，即在当前条件下，首先识别出影响力大的核心专利，也就是"高影响力专利"，并假定这些核心专利相对其他一般专利更有可能产生大的经济价值。

关于如何发现核心专利，国内外学者都进行了积极探索，产生了较多的研究成果，主要包括基于外部特征的核心专利识别、基于引用网络的核心专利识别、采用布拉德福定律[5-7]的核心专利识别以及利用专利分析软件[8-9]进行的核心专利识别，其中研究最多的主要集中在两方面。

第一是基于专利外部特征的核心专利发现研究。学者们主要选取了有代表性的被引次数、同族专利数和专利权利要求数等单一指标或组合指标来进行核心专利识别。F. Narin等[10-11]研究发现专利被引次数可以作为评估企业核心技术的指标，那些具有开创性的专利的被引次数均高于一般的专利。D. Harhoff等[12]则指出同族专利数在评估专利价值方面有很好的应用，F. Berger等[13]也发现核心专利的权利要求数多于普通专利的权利要求数，证明了权利要求数也可以用来识别核心专利。但是，通过单一指标来识别核心专

作者简介：张欣（ORCID：0000－0001－9697－0527），博士研究生；马瑞敏，副教授，博士，通讯作者，E-mail：ruimin.ma@sxu.edu.cn。
收稿日期：2017－09－25 修回日期：2018－01－28 本文起止页码：106－115 本文责任编辑：王善军

利具有很大的片面性和局限性,因此后来学者们提出利用指标体系来进行核心专利的识别。比如霍翠婷[14]选用了技术因素、法律因素、经济因素、企业因素和环境因素 5 个因素构建了企业核心专利识别的指标体系;谢萍、袁润和钱过[15-16]则是选取了专利发明人数、专利权人数、施引专利计数、同组专利数和权利要求数等 8 个指标来构建指标体系;李治东等[17]从专利申请、授权和无效三个阶段角度出发,选取了专利被引次数、同族专利大小、发明人数量和专利诉讼数量等指标来构建指标体系;王天歌等[18]从技术指标、经济指标和法律指标即被引次数、技术覆盖范围、权利要求数和专利有效性等指标出发构建了指标体系,并将其应用到生物医药领域中,这些指标体系的权重确定方法有专家打分法、TOPSIS 方法、熵权层次分析法和粗糙集理论方法,但是这些评价指标体系都存在一些共性问题,即指标体系和体系权重的确定或多或少都有一定的主观性,容易引起争议。

第二是基于专利引文的核心专利发现的探讨。其中最典型的就是利用专利共引来识别核心专利,这种方法一般先用表征专利外部特征的被引次数来初筛高价值专利,然后再利用筛选后的高价值专利建立共引矩阵求得相似度后进行因子分析或者聚类分析来达到识别核心专利的目的[19-21]。另外,还有学者利用潜在引用网络关系对专利价值进行了评估,例如冯岭等[22]通过计算专利间的相似性来建立专利潜在引用关联并计算直接和间接的被引次数来得到专利价值。然而,这些研究基本都是孤立地看待每一篇专利文献的引用情况,忽略了专利文献彼此之间的相互作用,一些作者也试图来解决这一问题。比如,S. Kim 等[23]对 PageRank 应用到专利引文网络之后的结果与原始的被引次数进行了比较分析,发现 PageRank 所得出的排名结果与被引次数有很强的相关性;顾立平[24]也将 PageRank 算法应用到美国专利引文网络中对专利进行了排名,发现 PageRank 算法确实优化了原始的被引次数排名,可见,PageRank 在改善原始的被引次数排名方面确实起到了不错的效果,但是这些学者在将 PageRank 算法应用到专利引文网络时并没有考虑到专利引文网络的特点,也没有考虑专利的时效性因素。

基于以上的综述,可以发现已有的研究存在这样的问题:一是基于外部特征的指标体系和其权重的确定具有主观性,二是运用引文网络对专利进行研究时,有少数学者运用了 PageRank 算法,考虑了专利之间的相互作用,但考虑专利的属性而对 PageRank 算法进一步改进来识别核心专利仍然是一大空白。因此,本文从加权引用的角度,考虑专利引文的特点,从引用的“权威性”和“时间性”两方面对原始的 PageRank 算法进行改进,提出一种新的 PatentRank 算法(简称 PTR),继而将其应用到 OLED 领域中来识别核心技术,验证 PTR 算法的科学性。

2　基于改进 PageRank 算法的核心专利发现模型构建

2.1　传统 PageRank 算法的基本原理

PageRank 算法是 Google 创始人 L. Page 和 S. Brin[25]于 1998 年构建早期的搜索系统原型时提出的链接分析算法,用来对互联网网页进行排名。对于某个网页 A 来说,其 PageRank 值基于两点假设:一是数量假设,在 WWW 网图模型中,一个网页收到的入链数越多,这个网页的质量就越高;二是质量假设,指向 A 页面的网页的质量参差不齐,质量高的网页会通过链接给其他页面传递更多的权重,因此,指向页面 A 的网页的质量越高,A 页面的质量也就越高。把互联网抽象成一个有向图模型,假设网页的个数为 n,该模型可以表示为 G = (V, E),其中 V 表示顶点,E 表示边,顶点的个数为 n,通过图的链接关系,可以建立邻接矩阵 H,任取 $h_{ij} \in H$,如果存在从网页 i 到 j 的链接,则 h_{ij} = 1,否则为 0,即

$h_{ij} = \begin{cases} 1, \text{存在从 i 到 j 的链接} \\ 0, \text{不存在 i 到 j 的链接} \end{cases}$,那么定义节点 i 的出度为 $O_i = \sum_{j=1}^{n} h_{ij}$, $i = 1,2,3,\cdots n$,从而节点 j 的 PageRank 值如公式(1)[25]所示:

$$PR(j) = \sum_{i=1}^{n} \frac{PR(i)}{O_i} \qquad \text{公式(1)}$$

即节点 j 的 PageRank 值不仅受节点 i 的 PageRank 值的影响,还受节点 i 的出度的影响,这是最初的模型。将邻接矩阵 H 矩阵中的每一个元素都除以每一行和,则得到归一化的邻接矩阵 A,即 $A_{ij} = \begin{cases} h_{ij}/O_i, & if\ O_i > 0 \\ 0, & if\ O_i = 0 \end{cases}$,这样,定义图 G 的转移概率矩阵为 M,将邻接矩阵 A 转置之后便得矩阵 M,矩阵 M 中任意一个元素 M_{ij} 表示从网页 j 到 i 的条件转移概率 P(i | j),这样每个节点的 PageRank 值就可以简写为 P = MP。很明显从 P = MP 这个式子中可以看出这是要求矩阵 M 的特征向量 P,而且该特征向量对应的特征值为 1,这可以用幂迭代法来实现。然而,由于网络中可能会存在“悬挂节点”,即出度为 0 的节点,也就是对应

转移概率矩阵中某一列为 0,导致该矩阵无法收敛,因此,后来就引入了阻尼因子 λ 来对公式(1)进行改进,即随机冲浪者以 λ 的概率沿着原始的链接关系进行游走,而以(1 - λ)的概率随机跳转到网络中的任何一个网页,这样改进后的网页 PageRank 值如公式(2)[26]所示:

$$PR(j) = \frac{1-\lambda}{n} + \lambda \sum_{i=1}^{n} PR(i)/O_i \quad 公式(2)$$

其中 λ 为阻尼因子,一般取值 0.85。将公式 2 简写成 $P = (\frac{(1-\lambda)E}{n} + \lambda M) * P$,其中 E 是 n * n 的全 1 矩阵,这里新的转移概率矩阵实质上就是 $B = \frac{(1-\lambda)E}{n} + \lambda M$。在开始时,赋予给每个网页一个初始的 PR 值,这个初始的 PR 值大小无关紧要,随着迭代循环次数的增多,每个网页的 PR 值最终会收敛到一个稳定的值,即最后的稳态分布列向量 PR,也就是转移概率矩阵 B 的特征向量,改进后的公式(2)是被应用最广的。

2.2 传统 PageRank 算法在核心专利发现的适用性

目前,有许多学者利用公式(2)或者对公式(2)进一步改进并将其应用到期刊、论文和作者引文网络中来评价期刊、论文和作者的影响力。例如在期刊影响力评价方面,马凤[27]利用原始的 PageRank 算法对图情领域的期刊进行了评价研究;文献评价方面,马楠和官建成[28]则将 PageRank 算法应用到分子生物学领域的文献中来发现重要的文献,段庆锋等[29]针对 PageRank 算法倾向于发表时间已久的文章的缺点,将引文间隔时间引入算法中,对原始 PageRank 算法进行了改进,优化了评价的结果;在作者影响力评价方面,E. Yan 等[30]将改进的 PageRank 算法应用到作者合作网络中来发现作者的影响力,并收到了不错的效果。可见,随着时间的推移,从不同角度考量不同因素的各种各样改进的 PageRank 算法不断产生并应用到评价实践中。

同样地,与期刊、论文和作者一样,专利文献里也存在丰富的引用关系,这称之为专利引文,这里主要采用基于审查员的引用。在专利引文中,根据节点的不同,一般将专利引用网络分为专利引文网络、专利权人引用网络、专利技术领域关联网络和专利引用学术文献网络等[31],本文研究的网络属于专利引文网络,即网络两边的节点是专利文献。网络中所有的节点代表专利,边代表专利之间的引用关系。专利引文网络和网页链接网络从图论的角度看均具有相同的拓扑结构,均是由节点和连接节点的边组成,节点代表个体,边代表关系,这两个网络在本质上是相似的,因此,将链接网络中的 PageRank 算法应用到专利引文网络中是可行的。但同时二者也有一些不同,首先,在网页链接关系中,彼此之间可以互相引用,引用是相互的,且不考虑时间的先后顺序,而在专利引用网络关系中,引用关系考虑先后顺序,只能是后发表的专利对早先发表的专利的引用,说明了专利引文网络对时序的要求;其次,专利引文网络是静态的,而链接网络是动态的;再次,专利引文网络的链接具有目的性和集中性,一般专利会倾向于链接与当前主题相关或者高影响力的专利,而链接网络中的链接具有随意性。因此将 PageRank 算法用于专利引文网络时需要做出相应的改进。

2.3 新模型的构建

正如前文所言,链接网络和专利引用网络之间存在一定的差别,其中最主要的差别是专利引用网络中引用的时间性和目的性。首先,原始的 PageRank 算法是针对静态网页的排序算法,只考虑网页的总被引次数,不考虑网页发表的时间,但是在专利网络中,假如存在专利 A 和专利 B,它们的发布时间分别为 2005 年和 2010 年,到目前为止,专利 A 和专利 B 的被引用次数一样,即 C(A) = C(B),但是很明显,单位时间内专利 B 的被引用次数明显高于专利 A,专利 B 的影响力是高于专利 A 的,造成这种现象的原因是专利文献的老化,即年代久远的专利并不会永远考虑被最新的专利所引用。戈斯内尔(C. F. Gosnell)[32]曾指出在知识的累积过程中,随着时间的推移,一切知识或其相应的载体会逐渐失去原有价值,提出用“文献老化”表示这种文献资料逐渐变得不再有用或不再有效的过程。专利作为知识的载体,也存在老化情况,即随着时间的增长,年代久远的专利在新公布专利引文中的被引次数占比在逐渐减少,但是原始的 PageRank 算法不考虑时间的因素,只考虑总被引次数,这样使得最新发表的专利其潜在价值就难以被发现,因此引入“时间性”因子。其次,专利引用并不是随意的,而是有很强的目的性,新发表的专利一般会引用与自己相关度高且权威性很高的基础专利,在社会网络中常常运用度来描述一个角色的影响力,度越高说明该角色越活跃,也越显著,所以如果一个专利的入度很高,那么该专利受到的认可度也越高,被引用的概率也就越大,因此,在专利引用网络中进行值传递时应该给予入度高的专利更大的权重,这里定义专利的入度中心度即为专利的权威性,引入“权威性”因子。

基于这两点,本文对传统的 PageRank 算法进行了改进:随机冲浪者在以 λ 的概率按照引用链接进行跳转时不是以等概率进行跳转,而要考虑被引专利的“权威性”和“时间性”。在转移概率矩阵 M 部分,将节点“权威性”和“时间性”按照一定的比例加权考虑进来,设比例性因子为 α。节点的“权威性”用 W_a 表示,节点的“时间性”采用文献老化的负指数模型,具体到每一个专利文献,其老化模型可采用贝尔纳的负指数模型[33]:$C(t) = C_0 e^{-bt}$,其中 t 表示专利的年龄,即统计年与发表年的差值,C(t)表示年龄为 t 的专利在统计年的被引次数,C_0 为常数,b 为专利文献的老化率,该部分用 W_t 表示,基于此,本文建立如下新模型,如公式(3)所示:

$$PTR(u) = (1-\lambda)/N + \lambda \sum_{v \in I(u)} PTR(v)[\alpha W'_a + (1-\alpha) W'_t] \quad \text{公式(3)}$$

其中 PTR(u)和 PTR(v)分别表示专利 u 和专利 v 的 PatentRank 值,λ 是阻尼因子,一般取 λ 为 0.85,N 是网络中节点的个数,也即专利数。α 是权威性的比例因子,则$(1-\alpha)$是时间性的比例因子。

$W_a' = \frac{W_a}{MaxW_a}$,$W_a$ 是专利 u 的入度,Max W_a 是专利 v 的出度的节点中的最大入度数,W_a'是专利 u 的入度归一化。

$$W'_t = \frac{W_t}{MaxW_t},\quad W_t(v,u) = \begin{cases} C_0 e^{-\beta(T_v - T_u)}, & h_{vu}=1 \\ 0, & h_{vu}=0 \end{cases},$$

$W_t(v,u)$是专利 u 和专利 v 的“时间性”因子,C_0 是常数,β 是老化指数,T_v 和 T_u 分别是专利 v 和专利 u 的发表时间,$(T_v - T_u)$即专利 v 和专利 u 的时间差,W'_t 是专利 u 和专利 v 的“时间性”因子的归一化,从公式可以看出,专利之间的时间差越大,其“时间性”因子值越小,符合本文的预设。

3 实证分析

OLED(Organic Light Emitting Diode,有机发光二极管)是一种新兴流行的电子显示技术,涉及技术主要有电致发光光源、电致发光材料和显示技术等,具有很好的应用前景,得到产业界的广泛关注。下面将通过对 OLED 领域的实证分析来检验本文所提方法的科学性。

3.1 数据来源与处理

专利数据来源于德温特(Derwent Innovations Index,DII)数据库,根据 OLED 领域的国际专利分类号以及关键词等进行联合检索,确定了检索策略为 TI = (“organic light emitting diode * ” or “organic light emitting display * ” or oled or oleds or pleds or pled or “p led” or “organic electroluminescent” or oel or oeld) or TS = (“small molecular organic light emitting” or smoled or smoleds or “sm oled” or “sm oleds” or “polymer organic light emitting” or “polymer light emitting diode * ” or “polymer light emitting display * ”) or (IP = (H05B－033 * or G09G－003 * or C09K－011 * or H01L－051 * or H01L－033 * or H01L－027 * or G09F－009 * or G02F－001 * or C23C－014 * or C07C－211 * or H01L－031 * or C07F－015 * or C08G－061 * or G01R－031 * or G09G－005 *) and TS = (“organic light emitting diode * ” or “organic light emitting display * ” or oled or oleds or pleds or pled or “p led” or oeld)),时间跨度为 1963 年－2016 年(搜集时间为 2016 年 10 月 27 日),最后共搜集下载得到 51 367 条专利数据。

对下载的数据进行进一步处理。首先,抽取专利之间的引用关系。利用 java 程序对专利文档中的 PN (Patent Number)字段和 CP(Cited Number)字段进行“引用－被引用关系”的抽取,并将同族专利利用其中一个专利来代替,形成一对一的专利引用关系。其次,抽取每个专利的公开日期。本文选择了专利文档中的 GA 字段进行时间的抽取,GA 字段是德温特主入藏号,它是德温特分配给每个专利族的第一个被其收录专利的唯一确认号,其前四位是年号,这样,通过 GA 字段得到了每个专利的公开时间,对于那些不在搜集范围内的专利,本文采用手动检索德温特数据库来获得其时间。在完成字段的抽取之后,将数据导入到 Pajek 软件中,通过抽取最大连通图,共得到 147 694 个专利节点,对这些数据进行“权威性”和“时间性”计算之后,利用 Matlab 软件进行 PTR 的计算。

3.2 参数 α 和 C_0、β 的确定

参数 α 是专利“权威性”的权重,本文利用熵权法来确定该权重。熵权法是一种客观赋权法,不受主观因素影响,在这里采用熵权法对“权威性”和“时间性”的权重进行分配,通过计算得到 α 为 0.9,即“权威性”因子的权重为 0.9,“时间性”因子的权重为 0.1。

前文中提到的 PTR 算法中的“时间性”因子服从负指数模型 $W_t(v,u) = \begin{cases} C_0 e^{-\beta(T_v - T_u)}, h_{vu}=1 \\ 0,\ h_{vu}=0 \end{cases}$,$\beta$ 是专利

文献的老化率,由于2016 年数据不全,这里以 2015 年为统计年绘制专利老化曲线,如图 1 所示,其中横坐标为专利年龄,纵坐标为在统计年各个年龄阶段专利的被引次数,并得到拟合曲线 $y = 893.8 * e^{-0.13x}$,拟合优度 $R^2 = 0.809$,说明拟合效果较好,从拟合曲线得知专利文献的老化率 β 为 0.13,C_0 为 893.8。

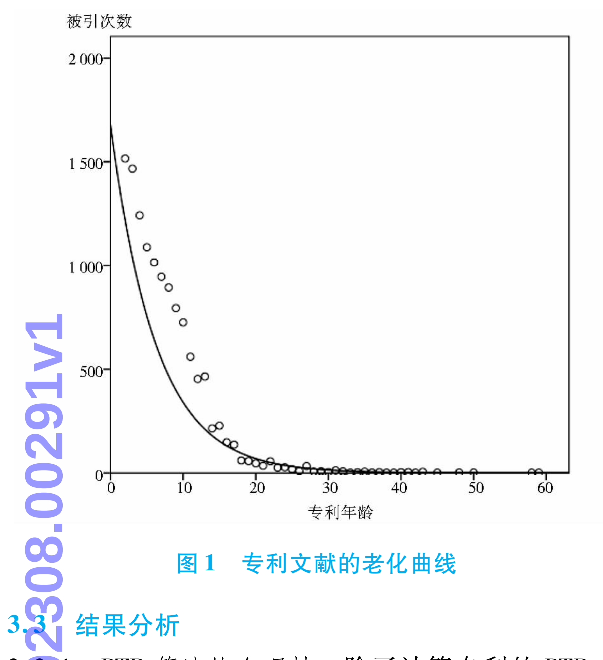


图 1 专利文献的老化曲线

3.3 结果分析

3.3.1 PTR 算法的合理性 除了计算专利的 PTR 值外,本文还对专利的被引次数和原始 PageRank 值(简称 PR 值)进行了计算,表 1 列出了 PTR 排名前 30 名专利的具体情况。从表 1 中可以看出,PTR 排名前 10 名专利中,PR 和被引次数排名也在前 10 名的专利均有 8 个,占 80%;在 PTR 排名前 20 名专利中,PR 和被引次数排名也在前 20 名的专利各有 18 个和 15 个,分别占 90% 和 75%,而在 PTR 排名前 30 名专利中,PR 和被引次数排名也在前 30 名的专利分别占 83.3% 和 66.7%,从该统计数据可以看出,PTR 算法靠前的专利一般也是被引次数和 PR 算法排名靠前的专利。从专利权人看,这些专利分别属于美国伊斯曼柯达公司、美国普林斯顿大学、英国剑桥显示技术公司、韩国三星电子、日本新日铁化工有限公司和日本先锋电子等公司,其中柯达公司在该领域属于领头羊,前 30 名专利中有 10 个专利都属于该公司,占到专利权人的 1/3,通过查看发明人可知,这些专利的发明人是该公司的邓青云(C. W. TANG)博士等一拨人,邓青云也因此被誉为 OLED 之父,在邓青云博士的带领下,柯达公司在 OLED 早期研发过程中占据主导地位,所以排名前几位专利都属于柯达公司。在专利权人中,排名第二位的是美国的普林斯顿大学,其发明的专利占到前 30 名的 1/6,其代表性发明人 S. Forrest 教授曾经因为在 OLED 领域的贡献获得美国 IPO 国家杰出发明奖和托马斯爱迪生奖,所以普林斯顿大学的 OLED 技术也是遥遥领先。除了 OLED 小分子材料的研究以外,英国剑桥显示技术公司(CDT)于 1990 年公布的 EP423283-A 专利(表 1 中排名第 10)首次研发出了以共轭高分子 PPV 为发光层的 OLED,而这一发明引发了采用高分子材料发光的 OLED 研究热潮,因此该专利在高分子 OLED 研究方面具有很强的开创性。日本的先锋公司则在显示屏技术方面有一定的研发优势,表中的 JP8315981-A 专利正是关于显示面板的专利。通过上述分析,可以发现除个别特殊情况之外,PTR 排名靠前的专利其 PR 排名和被引次数排名也很靠前,而且 PTR 排名靠前的专利均是领域中比较重要的专利,因此 PTR 算法可以将领域中一些关键专利识别出来。

一般来说,就某一问题所提出的改进算法,其运算结果不应该与已有算法所得出的结果有太大的偏差或者甚至颠覆原来的结果,否则这样的算法改进就被认为是不合理的。表 1 已经对前 30 名专利进行了分析,现在通过计算 PTR 算法与 PR、被引次数的 Spearman 等级相关性来进一步证明该算法的合理性。本文涉及的专利节点总共有 147 600 个,一方面节点太多,不方便把所有的节点都放到 SPSS 中进行计算,另一方面,排名靠后的节点其 PTR 和 PR 得分差距很小,对它们进行相关性分析没有实质意义。因此,这里截取总节点数的 1% 进行研究,即按照被引次数将节点从大到小降序排列,取前 1 476 个节点进行分段分析(第1 476 个节点处正好是被引次数为 20 的节点),分别以被引次数≥100、≥60、≥50、≥40、≥30 和≥20 为分界点进行分析,涉及的节点个数分别为 37、123、176、411、666 和 1 476 个,分析结果如表 2 所示。从表中可以看出,在被引次数大于 40 次时,被引次数与 PTR 的相关系数、被引次数与 PR 的相关系数均大于 0.5,且在被引次数大于 100 时,相关系数最大,均在 0.7 以上。同时,PTR 算法与被引次数的相关系数在各个分段都优于 PR 算法与被引次数的相关系数。另外,还可以看到原始的 PR 算法和改进后的 PTR 算法的相关系数一直很高,维持在 0.9 以上,说明 PTR 算法与 PR 算法在原理上有高度的一致性。

3.3.2 PTR 算法的优越性 虽然三种算法在排名上大致趋同,但是就各个专利而言,它们之间的排名还是有一定的差距,图 2 展示了 PTR 算法排名前 30 名专利

张欣，马瑞敏. 基于改进 PageRank 算法的核心专利发现研究［J］. 图书情报工作,2018,62(10):106－115.

表 1　PTR 排名前 30 名专利的具体排名情况

专利号	专利权人	年份	PTR 排名	PR 排名	被引次数排名
US4356429-A	EASTMAN KODAK	1982	1	3	4
US4539507-A	EASTMAN KODAK	1984	2	2	2
US4769292-A	EASTMAN KODAK	1988	3	1	1
US4720432-A	EASTMAN KODAK	1988	4	5	7
US4885211-A	EASTMAN KODAK	1988	5	4	10
US5811833-A	UNIV PRINCETON	1998	6	7	3
US3172862-A	DOW CHEMICAL CO	1965	7	15	68
US3173050-A	DOW CHEMICAL CO	1965	8	14	87
US6030715-A	UNIV PRINCETON	2000	9	9	5
EP423283-A	CAMBRIDGE DISPLAY TECHNOLOGY	1990	10	6	6
US3710167-A	RADIO CO of AMERICA	1973	11	17	88
US5703436-A	UNIV PRINCETON	1997	12	10	8
JP8315981-A	PIONEER ELECTRONIC CO	1996	13	8	17
US6229506-B1	SARNOFF CORP	1998	14	11	15
US5151629-A	EASTMAN KODAK	1992	15	12	11
US20060007072-A1	SAMSUNG ELECTRONICS CO	2006	16	19	14
US5061569-A	EASTMAN KODAK	1991	17	26	9
US5294870-A	EASTMAN KODAK	1993	18	13	19
JP10319908-A	SARNOFF CO	1999	19	16	30
US3995299-A	UK SEC INDUSTRY	1976	20	42	123
US3530325-A	AMERICAN CYANAMID	1970	21	21	105
US5707745-A	UNIV PRINCETON	1996	22	27	13
US4164431-A	EASTMAN KODAK	1979	23	406	127
EP855848-A2	INT MFG & ENG SERVICES CO LTD	1998	24	22	31
US6097147-A	UNIV PRINCETON	2000	25	37	12
US3621321-A	EASTMAN KODAK	1971	26	24	120
US4020389-A	MINNESOTA MINING CO	1977	27	415	125
WO2007063754-A1	NIPPON STEEL CHEM CO	2008	28	18	18
EP1061497-A1	SONY CO	2001	29	40	25
WO200041893-A1	3M INNOVATIVE PROPERTIES CO	2005	30	30	32

表 2　被引次数、PR 和 PTR 之间的 Spearman 相关系数

各分组 \ 被引次数	被引次数≥100	被引次数≥60	被引次数≥50	被引次数≥40	被引次数≥30	被引次数≥20
被引次数与 PTR	0.812	0.561	0.567	0.744	0.451	0.467
被引次数与 PR	0.761	0.505	0.507	0.706	0.399	0.426
PTR 与 PR	0.926	0.965	0.975	0.985	0.984	0.985

的三种算法的趋势走向情况。可以看到有 8 个特殊的专利,它们在三者排名上的差异较大(具体排名见表 3),其中 US3172862-A、US3173050-A、US3710167-A、US3995299-A、US3530325-A 和 US3621321-A 这 6 个专利 PTR 和 PR 排名差别很小,排名均在 50 名之内,但是被引次数排名却与 PTR 和 PR 排名相差过大,排名均在 50 名之外。另外两个特殊的专利 US4164431-A 和专利 US4020389-A 的被引次数和原始 PR 排名都很靠后,而 PTR 算法排名却很靠前,从图中可以看到这两个专利的被引次数排名在 125 名左右,PR 排名在 200 名之外,而 PTR 排名却很靠前,排在 25 名左右,差距悬殊。

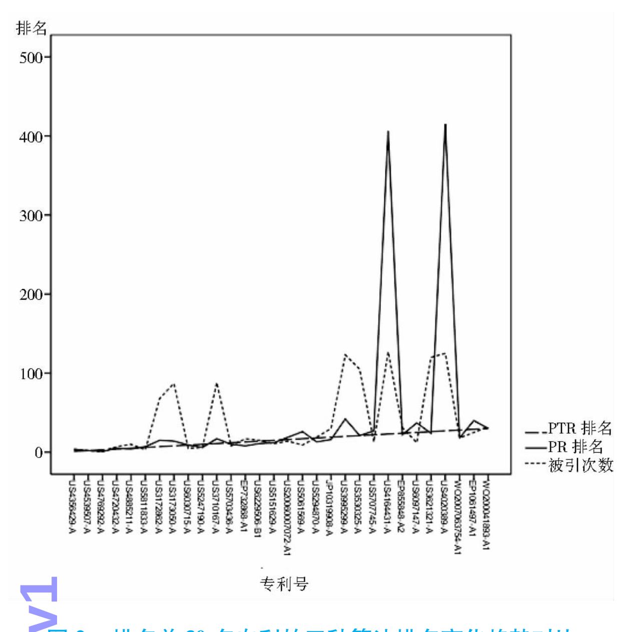


图 2　排名前 30 名专利的三种算法排名变化趋势对比

从表 1 这些特殊专利的专利权人中得知这些专利分别属于美国陶氏化学公司、美国无线电公司、英国 SEC 行业、美国 Cyanamid 公司、美国柯达、美国 3M 公司，这些公司都是 90 年代初发展起来的大公司，专利涉及电致发光电池、电致发光构造和电致发光材料等，发表时间在 20 世纪 60 和 70 年代左右。从表 3 中发现除了 US3172862-A、US3621321-A 和 US4020389-A 专利的权利要求数在 6 个左右外，其余专利的权利要求数均在 10 个以上，在专利审核过程中，权利要求数不超过 10 个时是不收费的，但当权利要求数超过 10 个时，每条专利会收取一定的附加费，所以一般认为权利要求数超过 10 条的专利均可视为比较重要的专利[34]，也就说这些特殊的专利均是很重要的专利。从这些专利的施引专利中发现这些专利均是 PTR 算法中排名靠前的专利，大部分排名在前 5 名左右，说明这些发表

表 3　8 个特殊专利的施引专利以及该专利的相关排名情况

施引专利	被引专利	发表时间	PTR 排名	PR 排名	被引次数排名	权利要求数	平均被引时间间隔
US4720432-A	US3172862-A	1965	7	15	68	6	39
US4356429-A							
US4769292-A							
US4885211-A							
US4720432-A	US3173050-A	1965	8	14	87	13	38
US4356429-A							
US4539507-A							
US4769292-A							
US4885211-A							
US4720432-A	US3710167-A	1973	11	17	88	14	30
US4356429-A							
US4769292-A							
US4885211-A							
US4356429-A	US3995299-A	1976	20	42	123	13	17
US4539507-A							
US4769292-A							
US4885211-A							
US4356429-A	US3530325-A	1970	21	21	105	12	29
US4769292-A							
US4885211-A							
US4356429-A	US4164431-A	1979	23	406	127	17	23
US4356429-A	US3621321-A	1971	26	24	120	6	22
US4539507-A							
US4769292-A							
US4885211-A							
WO9733296-A1	US4020389-A	1977	27	415	125	7	27
US4356429-A							

在 60－70 年代的特殊专利为那些重要专利提供了技术支持,而且由于它们被高质量专利所引用,其 PTR 和 PR 排名相比被引次数排名前进了许多,这也解释了为什么图 2 中那 6 个特殊专利的被引次数排名特别靠后而 PTR 和 PR 排名却很靠前,这说明了 PTR 算法相比被引次数的优越性,其不仅可以识别出被引次数高的专利,而且还能识别出被高质量专利所引用的基础性专利。

接下来说明 PTR 算法相比 PR 算法的优越性,这里绘制了排名前 30 名专利的 PTR 值和 PR 值得分对比图。从图 3 中可以看到,整体上 PTR 得分要高于 PR 得分,排名在前 10 名左右专利的 PTR 和 PR 值差别较大,随着排名的推后,两条曲线趋于重合,看到二者在得分方面,PTR 算法相比 PR 算法有更大的区分度。从图 2 中得知,专利 US4164431-A 和专利 US4020389-A 的 PR 排名比 PTR 排名靠后很多,查看这两个专利的施引专利,发现均有专利 US4356429-A,该专利的 PR 算法排名在第三,而 PTR 算法排名在第一,这在一方面会影响其排名的变动,另一方面,因为 PTR 算法相较 PR 算法有较高的值区分度,尤其是排在前几位的专利,所以被 PTR 排名在第一位的 US4356429-A 专利所引用的那两个特殊专利的 PTR 排名相比 PR 排名前进了很多,这说明"值区分度"对引用了高影响力专利的那些专利有一定的影响力。

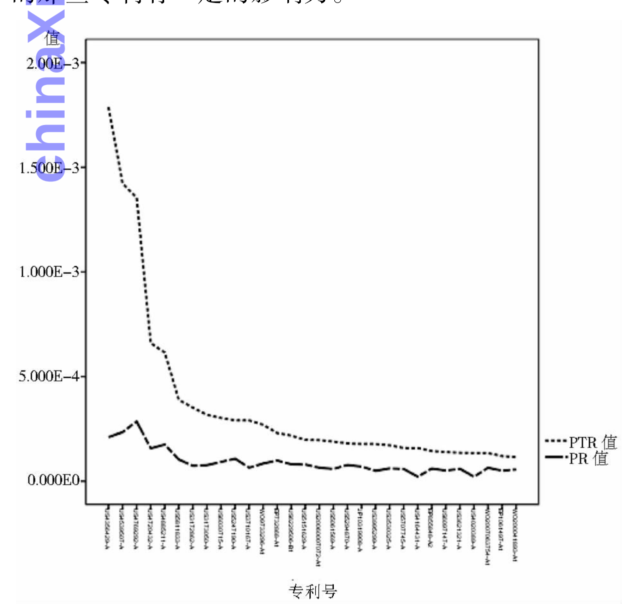


图 3　排名前 30 名专利的 PTR 和 PR 值得分对比图

PTR 与 PR 排名的差别是由算法改进部分被引次数和专利年龄引起的,这是一种综合的结果,因此只能在相对层面上来说明新改进算法的好处。首先,本文计算了被引次数和(PTR-PR)值之间的斯皮尔曼相关系数,将被引次数按照≥200 次、100－200 次和 50－100 次三个分段进行划分,查看其与(PTR-PR)值之间的关系,结果显示,在被引次数≥200 次时,被引次数与(PTR-PR)值之间的相关系数为 0.682,当被引次数在 100－200 次之间时,相关系数为 0.521,当被引次数在 50－100 次之间时,相关系数为 0.272,说明"权威性"因子对 PTR 值的影响会随着被引次数的增大而增大,从表 1 中可以看到,相同年份下,被引次数排名靠前的专利其 PTR 排名也相对靠前,例如发表于 1988 年的三个专利 PTR 排名大小顺序基本与被引次数相同。其次,这里专利年龄用专利的平均被引时间间隔来表示,平均被引时间间隔指一件专利从发明到被其他专利所引用的所有时间的平均,本文探究了专利平均被引时间间隔与(PTR-PR)值之间的关系。结果发现除个别特殊点之外,平均被引时间间隔对(PTR-PR)值的影响几乎为 0,即使有一些影响,也是在局部有微小的变动,例如在平均被引时间间隔位于 5－20 年之间的专利,其(PTR-PR)值有些变动,说明了"时间性"因子对 PTR 算法整体得分影响较小,从表 3 列出的这些专利的平均被引时间间隔可以看出,排名在前面的专利的被引次数也排在前面,但是平均被引时间间隔却不是很小,而对于被引次数排名差不多的专利,平均被引时间间隔小的专利排名就稍微有点优势,例如表 3 中的 US3995299-A 专利相比专利 US3621321-A 来说,被引次数排名不如 US3621321-A,但是平均被引时间间隔比 US3621321-A 小,所以 PTR 排名靠前,US4164431-A 和 US4020389-A 专利相比也是如此。因此,在针对专利引用网络进行研究时,专利的年龄对 PTR 的影响要远远小于被引次数,改进后的 PTR 在局部改善了 PR 的排名。

4　结论

本文首先界定了核心专利是"高影响力"专利的概念,然后总结了现有识别核心专利的方法,并抓住运用改进 PageRank 算法进行核心专利识别研究的空白领域进行研究。通过介绍原始 PageRank 算法的基本原理,并结合专利的两个属性"权威性"因子和"时间性"因子,对 PageRank 算法进行了改进,提出了改进后的 PTR 算法,也就是加权 PageRank 算法,并将其应用到 OLED 领域进行了实证分析,通过数据分析,可以得到以下几个结论:

(1)改进后的 PTR 算法不仅考虑了引用链接的次

数,还将专利的年龄和专利的被引次数考虑到了 PageRank 算法中,PageRank 算法本身和这些新增的加权条件均属于围绕专利的客观因素,因而主观性很弱,这使得评价结果更加科学。

(2)改进的 PTR 算法同被引次数、原始 PageRank 算法一样,能够识别出 OLED 领域的一些关键性专利。PTR 算法、PR 算法和被引次数之间的 Spearman 相关系数验证了被引次数与 PR 和 PTR 之间、PR 与 PTR 之间有较强的相关性,说明了改进后的 PTR 算法在识别核心专利方面与 PR 算法和被引次数的高度一致性,而且 PTR 算法排名前 30 名的专利基本上都是领域内比较重要的专利,说明用 PTR 算法来发现领域内的核心专利具有一定的合理性。

(3)改进后的 PTR 算法不仅能识别出被引次数高的核心专利,而且还能识别出被引次数低的基础性专利,体现了 PTR 算法相较被引次数的优越性。前文中提到的 8 个特殊的专利,其被引次数虽然较低,但是均受到高质量专利的认可,从而得到了较高的 PTR 值,发现这些基础性专利为研究该领域提供了追本溯源的研究思路,这也是 PTR 算法的优越性之一。

(4)改进后的 PTR 算法结果和 PR 算法结果高度吻合却又有所区别,PTR 算法相较 PR 算法具有更高的区分度。PTR 算法和 PR 算法的相关系数在 0.9 以上,说明二者在原理上的一致性,但是,从 PTR 排名前 30 名专利的二者得分曲线图又可以看到二者在表现好的专利之间,PTR 曲线要比 PR 曲线陡峭,其得分差距明显,说明了 PTR 算法相较 PR 算法可以将质量好的专利突显出来,区分度较高。

(5)改进后的 PTR 算法排名结果在局部调整了 PR 的排名,且 PTR 受到被引次数的影响要大于专利的年龄对 PTR 的影响,说明了在专利引用网络中,时间属性对网络中节点的影响较小,在之后的研究中可以忽略时间对专利引用网络的影响。

同时,本文在研究过程中也存在一些不足,首先,在进行 OLED 领域分析时,没有对其进行行业细分;其次,在 OLED 领域中,没有固定的“金标准”可以将被引次数和 PR、PTR 排名进行比较,在接下来的研究中,会选择相对比较成熟的具有公认的核心技术的领域进行研究,以便更科学地探究所要研究的方法。

参考文献:

[1]《国家知识产权战略纲要》(全文)[EB/OL].[2016-07-25]. http://www. sipo. gov. cn/ztzl/ywzt/zlwzn/xgljt/201306/t20130604_801744. html.
[2] 李建蓉. 专利文献与信息[M]. 北京:知识产权出版社,2002.
[3] 李建蓉. 专利信息与利用[M]. 北京:知识产权出版社,2006.
[4] 韩志华. 核心专利判别的综合指标体系研究[J]. 中国外资, 2010(4):193-196.
[5] 罗爱静, 尹瑾. 基于信息分析的中药领域核心专利技术发展研究[J]. 情报杂志, 2009, 28(S1): 37-39.
[6] 尹瑾, 赵玉梅. 我国中药核心专利技术区域分布研究[J]. 湖南工业大学学报(社会科学版), 2011, 16(3): 11-14.
[7] 胡晨希, 邵蓉. 基于布拉德福定律的药品核心专利分析[J]. 中国药事, 2012, 26(2): 134-136.
[8] 余敏杰, 田稷. 海洋生物产业专利情报分析[J]. 情报杂志, 2012(9): 11-14.
[9] 陆萍, 柯岚馨. Innography 在学科核心专利挖掘中的应用研究[J]. 图书馆工作与研究, 2012(8): 122-125.
[10] ALBERT M B, AVERY D, NARIN F, et al. Direct validation of citation counts as indicators of industrially important patents[J]. Research policy, 1991, 20(3):251-259.
[11] HARHOFF D, NARIN F, SCHERER F M, et al. Citation frequency and the value of patented inventions[J]. Review of Economics and statistics, 1999, 81(3):511-515.
[12] HARHOFF D, SCHERER F M, VOPEL K. Citations, family size, opposition and the value of patent rights[J]. Research policy, 2003, 32(8):1343-1363.
[13] BERGER F, BLIND K, THUMM N. Filing behaviour regarding essential patents in industry standards[J]. Research policy, 2012, 41(1):216-225.
[14] 霍翠婷. 企业核心专利判定的方法研究[J]. 情报杂志, 2012(11):95-99.
[15] 谢萍, 袁润, 钱过. 基于 TOPSIS 方法的核心专利识别研究[J]. 情报理论与实践, 2015, 38(6):88-92.
[16] 谢萍, 钱过, 袁润. 基于粗糙集理论的核心专利识别研究[J]. 情报杂志, 2015(7):34-38.
[17] 李治东,熊焰,方曦. 基于熵权层次分析法的核心专利识别应用研究[J]. 情报学报, 2016, 35(10):1101-1109.
[18] 王天歌,王金苗,袁红梅. 基于专利维度的我国生物医药核心技术的识别与分析[J]. 情报杂志, 2016, 35(4):112-117.
[19] WANG X, DUAN Y. Identifying core technology structure of electric vehicle industry through patent co-citation information[J]. Energy procedia, 2011, 5(1):2581-2585.
[20] WU H C, CHEN H Y, LEE K Y. Unveiling the core technology structure for companies through patent information[J]. Technological forecasting & social change, 2010, 77(7):1167-1178.
[21] LAI K K, WU S J. Using the patent co-citation approach to establish a new patent classification system[J]. Information processing & management, 2005, 41(2):313-330.
[22] 冯岭, 彭智勇, 刘斌, 等. 一种基于潜在引用网络的专利价值评估方法[J]. 计算机研究与发展, 2015, 52(3):649-660.
[23] KIM S, KIM S Y, HONG S W, et al. Applying PageRank to patent[C]// International conference on convergence content. Switz-

erland: Advanced Engineering Forum, 2013:7－8.
[24] 顾立平. 专利排名算法－运用引用次数与引文网络计算美国专利的研究[J]. 现代图书情报技术, 2011, 27(6):14－19.
[25] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer networks, 1998, 56(18):3825－3833.
[26] PAGE L. The PageRank citation ranking: bringing order to the Web, online manuscript[J]. Stanford digital libraries working paper,1998,9(1) :1－14.
[27] 马凤. 基于 PageRank 算法的期刊影响力研究[J]. 情报杂志, 2014(12):103－108.
[28] 马楠, 官建成. 基于网络结构挖掘算法的引文网络研究[J]. 情报学报,2008, 27(4):584－590.
[29] 段庆锋, 朱东华, 汪雪锋. 基于改进 PageRank 算法的引文文献排序方法[J]. 情报理论与实践, 2012, 35(1):115－119.
[30] YAN E, DING Y. Discovering author impact: a PageRank perspective[J]. Information processing & management, 2010, 47(1):125－134.
[31] 曹德斌. 专利引文网络分类分析及结构模式发现研究[D]. 长沙:国防科学技术大学,2013.
[32] GOSNELL C F. The rate of obsolescence in college library book collections as determined by an analysis of the three select lists of books for college libraries[D]. New York: New York University, 1943.
[33] 邱均平. 信息计量学[M]. 武汉:武汉大学出版社, 2007.
[34] 万里鹏. 我国企业专利权质押实证研究[D]. 重庆:西南政法大学, 2012.

作者贡献说明：

张欣:提出了改进的算法,完成论文撰写;
马瑞敏:对本文算法进行程序设计,并对论文进行整体上的把握与修改。

Research on the Discovery of Core Patents Based on Improved PageRank Algorithm

Zhang Xin[1,2]　Ma Ruimin[3]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190
[2] School of Ecnomics and Management, University of Chinese Academy of Sciences, Beijing 100190
[3] College of Economic and Management, Shanxi university, Taiyuan 030006

Abstract: [**Purpose/significance**] The discovery of core patents is an important part in technological innovation, which is of great significance to the technological improvement and the patent strategy layout. [**Method/process**] This paper firstly defined the concept of core patent. Then, by introducing the original PageRank algorithm model, the improved PageRank—— PatentRank (referred to as PTR) was put forward, which combining the cited times of patents and the patent age. Finally, we applied the new model to the field of OLED to identify the core patents. [**Result/conclusion**] We find that compared with the cited times, PTR can not only identify the core patents but also some basic important patents which provide the research ideas for the research of OLED technology. Besides, compared with the original PageRank algorithm, the values of PTR have higher degree of differentiation, and it can improve the PR results on the local scale.

Keywords: PageRank　core patents　OLED

下　期　要　目